Lec : 6

[2] Classification : discrete valued output.

① Binary Classification.

Has two different classes.

ex: Email : Spam / Not Spam

Tumor : Malignant / Benign.

y : output $\in \{0, 1\}$ binary classification problem

we can say $y \in \{1, 2\}$ is binary classification problem

Here, binary means 2 different classes.

0 : "Negative Class" (e.g benign tumor, spam email)

1 : "Positive Class" (e.g malignant tumor, not spam email)

② Multiclass Classification.

Has more than 2 different classes.

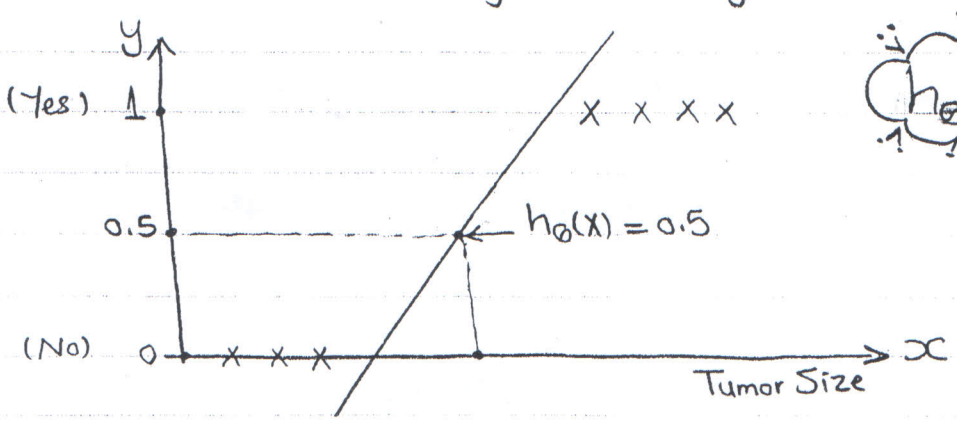
Y is the predicted variable.

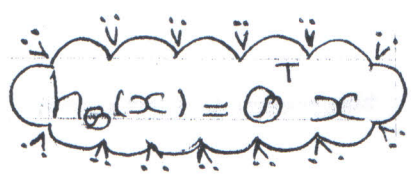
$Y \in \{0, 1, 2, 3\}$ is a multiclass classification Problem.

- <u>Using Linear Regression to get good Fitting to the data set.</u>

we explain by example

ex : Tumor : Malignant / Benign.





In this example <u>with this data set</u> we can get good filling to the data

1. We determine the threshold classifier output

→ Here, Threshold classifier output $h_\theta(x)$ at 0.5

. Note: 1. Threshold classifier outpout is very important to know the class in which every point (data set) belong

2. This point can take any value (not necessarily mean value القيمة المتوسطة)

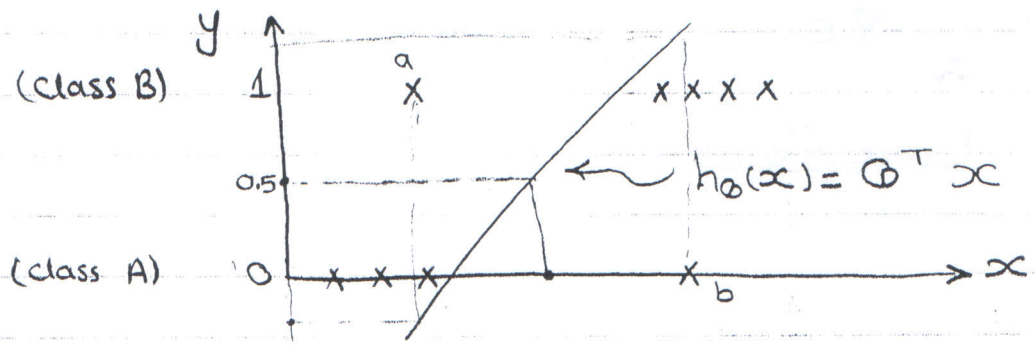
. We choose the value that give us the best Fitting.

→ From the threshold classifier output $h_\theta(x)$ at 0.5 we divide the data sets into two classes depending on the value of $h_\theta(x)$:

IF $h_\theta(x) \geq 0.5$, predict "$y = 1$"

IF $h_\theta(x) < 0.5$, predict "$y = 0$"

→ IF we have another data set



→ For a $h_\theta(x) < 0.5$ this mean that it belongs to class A but it belongs to class B in Fact.

→ For b $h_\theta(x) \geq 0.5$ this mean that it belongs to class B but it belongs to class A in Fact.

Note :. Using Linear regression, the hypothesis can output values much Longer than one or less than zero even if all of good the training examples have Labels y equals zero or one

. Generally, Linear Regression can't give good Fitting to the data set in all cases.

Logistic Regression

→ We have a problem using Linear regression when we get values of $h_\theta(x) > 1$ and values of $h_\theta(x) < 0$

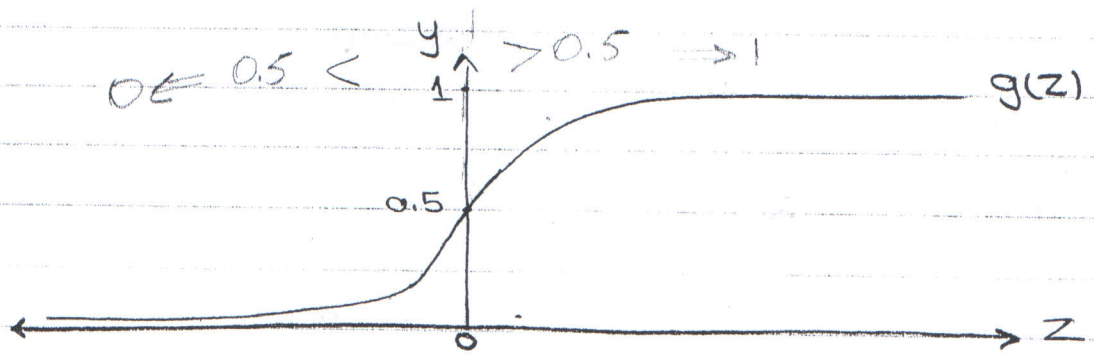
→ Want $0 \leq h_\theta(x) \leq 1$

$h_\theta(x) = g(\theta^T x)$

Where Linear regression model $h_\theta(x) = \theta^T x$

$$g(z) = \frac{1}{1 + e^{-z}}$$

Where $g(z)$: Sigmoid Function OR Logistic Function.



at $z \longrightarrow \infty$ $\quad g(z) = 1$

at $z \longrightarrow -\infty$ $\quad g(z) = 0$

Interpretation Of Hypothesis Output

$h_\theta(x)$ = estimated probability that $y=1$ on input x.

$$0 \leq h_\theta(x) \leq 1$$

probability of the input being belongs to one of 2 classes.

ex: IF $x = \begin{bmatrix} x_0 \\ x_1 \end{bmatrix} = \begin{bmatrix} 1 \\ \text{tumorSize} \end{bmatrix}$

IF we get $h_\theta(x) = 0.7$

this mean that 70% chance of tumor being malignant.

$h_\theta(x) = P(y=1 \mid x ; \theta)$

"probability that $y=1$, given x, parameterized by θ"

∴ $y=0$ or $y=1$

∴ $P(y=0 \mid x;\theta) + P(y=1 \mid x;\theta) = 1$

OR

$P(y=0 \mid x;\theta) = 1 - P(y=1 \mid x;\theta)$

* ——— *

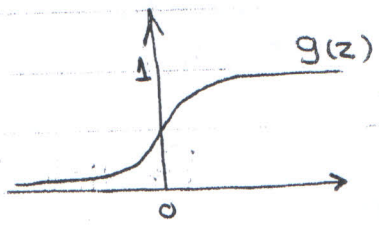
• Decision boundary

→ $h_\theta(x) = g(\theta^T x)$, $g(z) = \frac{1}{1+e^{-z}}$

→ Where $\theta^T x$: Equation.

• We will choose the best equation that give us the best Fitting

Predict "$y=1$" IF $h_\theta(x) \geq 0.5$



- $g(z) \geq 0.5$ when $z \geq 0$
- $h_\theta(x) \geq 0.5$ when $x \geq 0$
- $\because h_\theta(x) = g(\theta^T x)$
- $g(\theta^T x) = h_\theta(x) \geq 0.5$ whenever $\theta^T x \geq 0$

Predict "$y=0$" IF $h_\theta(x) < 0.5$

- $g(z) < 0.5$ when $z < 0$
- $g(\theta^T x) = h_\theta(x) < 0.5$ whenever $\theta^T x < 0$

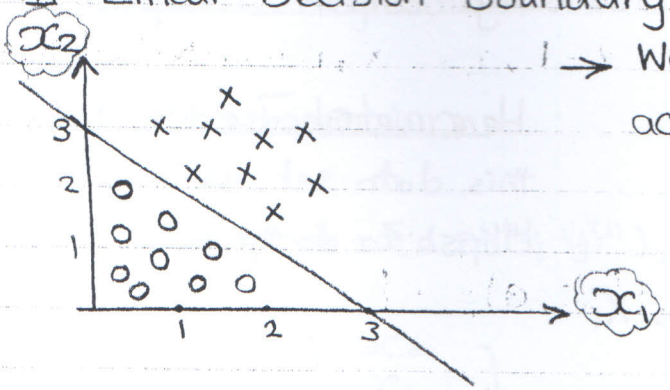
So, we can get the equation $\theta^T x$ and Apply the two conditions

$y=1$ ⟶ $\theta^T x \geq 0$

$y=1$ ⟶ $\theta^T x < 0$

we can replace the $\theta^T x$ by the paramters of the equation.

II Linear Decision Boundary



→ We choose the equation of $\theta^T x$ according to the data set

→ According to this data set we choose the Straight Line to Fit it, so equation of $\theta^T X$ is 1st order equation.

$$h_\theta(x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2)$$

$$\therefore \theta^T X = \theta_0 + \theta_1 x_1 + \theta_2 x_2 \quad = x_1 + x_2 = 3$$

Assuming $\theta_0 = -3$, $\theta_1 = \theta_2 = 1$

→ Predict "$y=1$" IF $\theta^T X \geq 0$

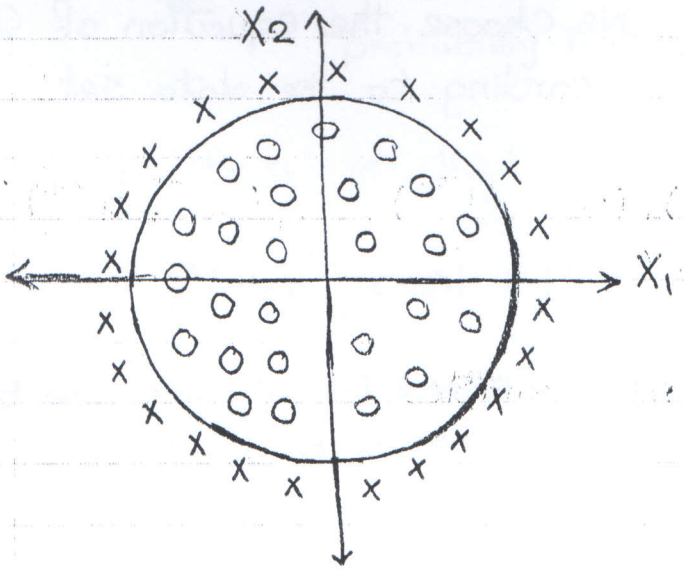
where $y=1$ when $g(\theta^T x) = h_\theta(x) \geq 0.5 \implies \theta^T X \geq 0$

Predict '$y=1$' IF $-3 + x_1 + x_2 \geq 0$

This mean for any example of features x_1 and x_2 that satisfy this equation $-3 + x_1 + x_2 \geq 0 \implies y=1$

$\therefore x_1 + x_2 = 3$ is the decision boundary where $h_\theta(x) = 0.5$.

[2] Non. Linear Decision Boundary.



Here we choose circle to fit this data set, we also can use ellipso to do the same thing

Polynomial Regression: add extra higher order polynomial Features for the hypothesis.

$$h_\theta(x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_1^2 + \theta_4 x_2^2)$$

Assuming $\theta_0 = -1$, $\theta_1 = \theta_2 = 0$, $\theta_3 = \theta_4 = 1$ علشان اوصل لمعادلة الدائرة

$\therefore$ Predict "$y=1$" IF $-1 + x_1^2 + x_2^2 \geq 0$

In another data set we add extra higher order polynomial

$$h_\theta(x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_1^2 + \theta_4 x_2^2 + \theta_5 x_1^3 \dots)$$

• Logistic Regression

J ← **Cost Function**

training set : $\{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \ldots, (x^{(m)}, y^{(m)})\}$

$$x \in \begin{bmatrix} x_0 \\ x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \quad m \text{ examples}$$

$x_0 = 1$

$y \in \{0, 1\}$

$$h_\theta(x) = \frac{1}{1 + e^{-\theta^T x}} \rightarrow z$$

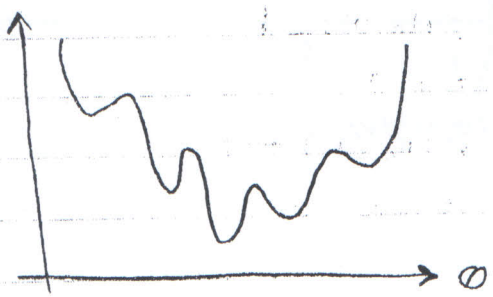
→ at Ideal Case Cost Function = Zero

- $h_\theta(x) = y = 1$ For data set $\in$ class A
- $h_\theta(x) = y = 0$ for data set $\in$ class B.

→ We choose parameters θ given this training sets which give us small value of Cost Function to get the best model fitting this data set.

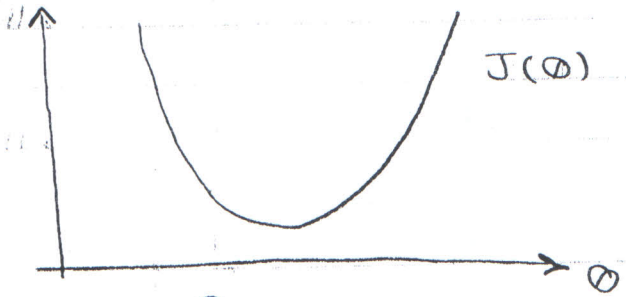
Linear regression: $J(\theta) = \frac{1}{2m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})^2$

Logistic regression: $\text{Cost}\,(h_\theta(x^{(i)}), y^{(i)}) = \frac{1}{2} (h_\theta(x^{(i)}) - y^{(i)})^2$



Not. Convex.

- more than one Local minimam

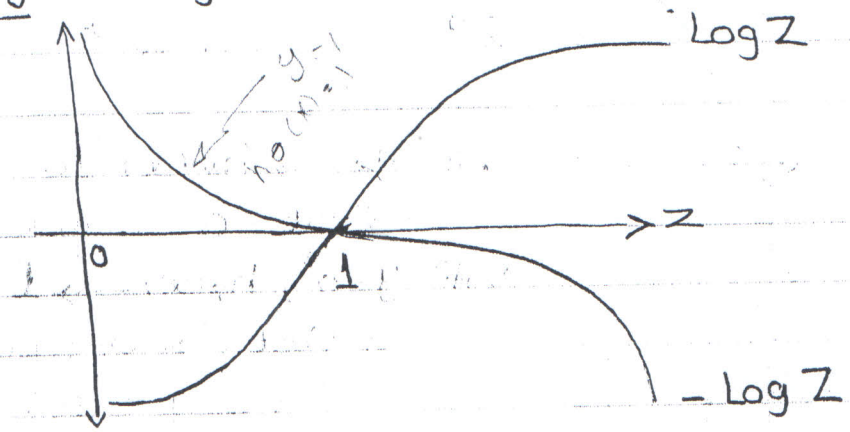


Convex

- one global minimum

→ We choose Convex Cost Function when using Logistic regression, as it has one global minimum. (this help us to get the smallest Cost Function value to get the best model fitting our data set.)

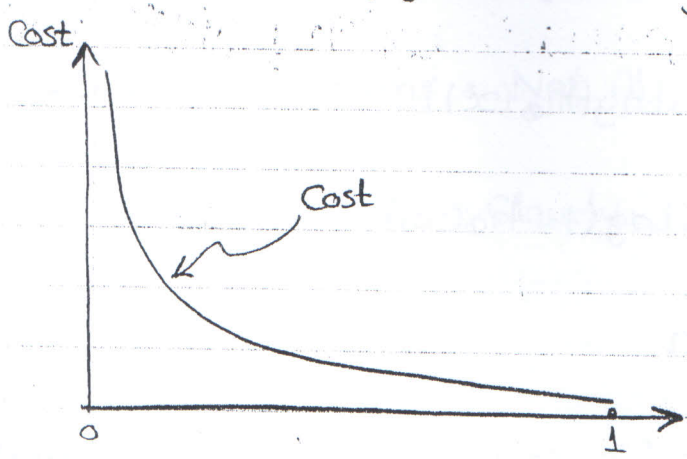
Logistic regression Cost Function.



$$\text{Cost}\,(h_\theta(x), y) = \begin{cases} -\text{Log}\,(h_\theta(x)) & y = 1 \\ -\text{Log}\,(1 - h_\theta(x)) & y = 0 \end{cases}$$

IF $Y = 1$

$$\text{Cost}(h_\theta(x), y) = -\log(h_\theta(x))$$



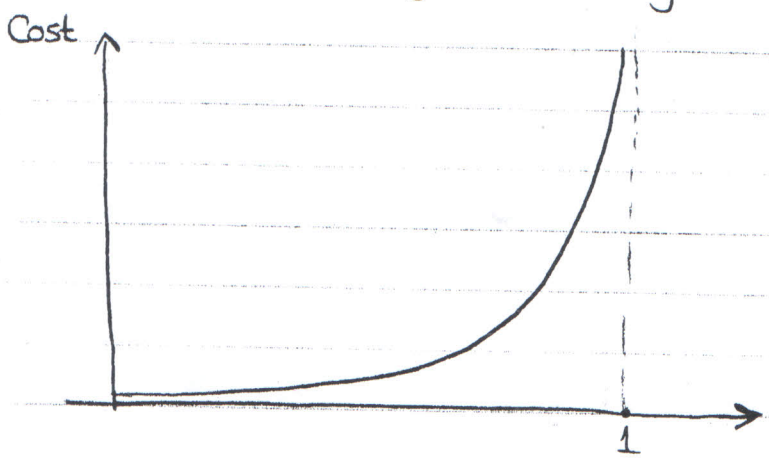
→ Ideal Case $y = h_\theta(x) = 1$

- IF $y=1$, $h_\theta(x)=1$
  $\therefore$ Cost $= 0$ (No Cost)
- IF $y=1$, $h_\theta(x) \to 0$
  $\therefore$ Cost $\to \infty$ (warse Case)
  here the point belongs to another class.

→ Captures intuition that if $h_\theta(x) = 0$ (predict $P(y=1|x;\theta) = 0$) but $y=1$, We will penalize Learning algorithm by a very Large Cost.

IF $Y = 0$

$$\text{Cost}(h_\theta(x), y) = -\log(1 - h_\theta(x))$$



→ Ideal Case $y = h_\theta(x) = 0$

- IF $y=0$, $h_\theta(x)=0$
  $\therefore$ Cost $= 0$ (No Cast)
- IF $y=0$, $h_\theta(x) \to 1$
  $\therefore$ Cost $\to \infty$ (worse case)
  (Large Cost)

Logistic regression Cost Function

$$J(\theta) = \frac{1}{m} \sum_{i=1}^{m} \text{Cost}\left(h_\theta(x^{(i)}), y^{(i)}\right)$$

$$\text{Cost}\left(h_\theta(x), y\right) = \begin{cases} -\text{Log}(h_\theta(x)) & y = 1 \\ -\text{Log}(1 - h_\theta(x)) & y = 0 \end{cases}$$

Simplified Cost Function

$$\rightarrow \text{Cost}\left(h_\theta(x), y\right) = -y \,\text{Log}\left(h_\theta(x)\right) - (1-y)\,\text{Log}\left(1 - h_\theta(x)\right)$$

at $y = 1 \rightarrow \text{Cost} = -\text{Log}(h_\theta(x))$

at $y = 0 \rightarrow \text{Cost} = -\text{Log}(1 - h_\theta(x))$

$$\rightarrow J(\theta) = -\frac{1}{m} \sum_{i=1}^{m} y^{(i)} \,\text{Log}\, h_\theta(x^{(i)}) + (1 - y^{(i)}) \,\text{Log}(1 - h_\theta(x^{(i)}))$$

$\rightarrow$ To Fit parameters θ : $\min J(\theta)$

$\rightarrow$ To make a prediction given new x:

$$\text{Output } h_\theta(x) = \frac{1}{1 + e^{-\theta^T x}}$$

$\rightarrow$ Gradient Descent

Repeat { $\theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta)$ }

Repeat {

$$\theta_j := \theta_j - \alpha \frac{1}{m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)}\right) x_j^{(i)}$$

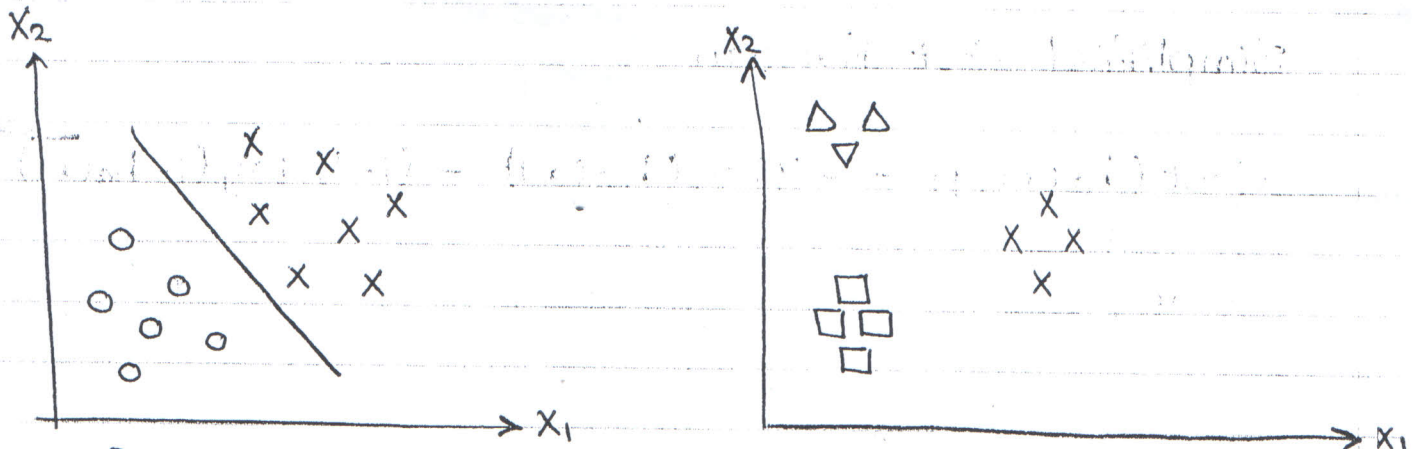
} (simultaneously update all θ_j)

Multiclass Classification

Email Folding / tagging : Work, Frinds, Family, Hobby.

Medical diagrams : Not ill, Cold, Flu.

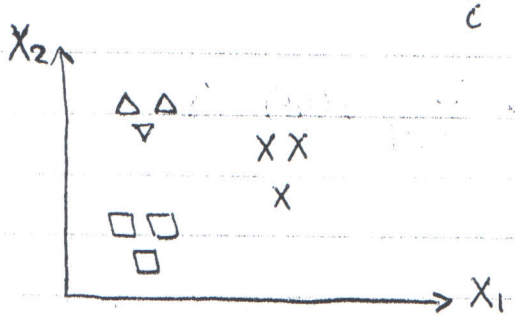
Weather : Sunny, Cloudy, Rain, Snow

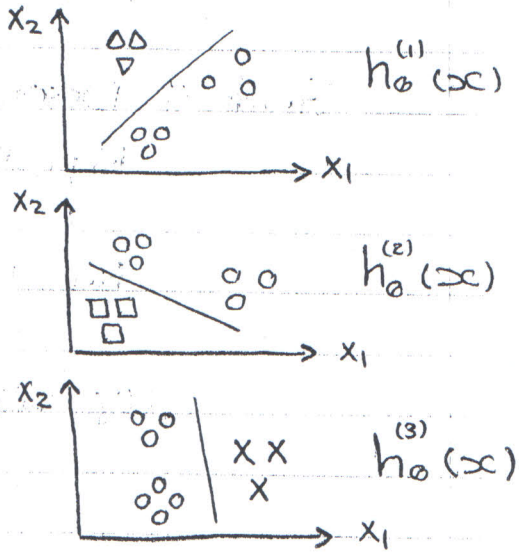


One-VS-all

- Train a Logistic regression classification $h_\theta^{(i)}(x)$ for each class i to predict the probability that $y = i$.

- On a new input x, to make a prediction, pick the class i that maximizes $\max_i h_\theta^{(i)}(x)$.





Class 1 : △

Class 2 : □

Class 3 : X

$h_\theta^{(i)}(x) = P(y = i | x; \theta) \quad (i = 1, 2, 3)$